Sequence divergence, functional constraint and selection in protein evolution

Justin C. Fay *
Department of Genome Sciences
Lawrence Berkeley National Laboratory
Berkeley, CA 94720
and
Chung-I Wu
Department of Ecology and Evolution
University of Chicago
Chicago, IL 60637

January 29, 2003

Abstract

The genome sequences of multiple species has made possible functional inferences from comparative genomics. A primary objective is to infer biological functions from the conservation of homologous DNA sequences between species. A second more difficult objective is to understand what functional DNA sequences have changed over time and are responsible for species' phenotypic differences. The neutral theory of molecular evolution provides a theoretical framework in which both objectives can be explicitly tested. Development of statistical tests within this framework has provided much insight into the evolutionary forces that constrain and in some cases change DNA sequences and the resulting patterns that emerge. Here, we review recent work

*Corresponding Author: Justin Fay (jcfay@lbl.gov), One Cyclotron Road, Mailstop 84-171, LBNL, Berkeley, CA 94720, Phone 510-486-6791. Fax 510-486-5614

on how functional constraint and changes in protein function are inferred from protein polymorphism and divergence data. We relate these studies to our understanding of the neutral theory and adaptive evolution.

Contents

1 Introduction

Evolution has left us with a fascinating puzzle. What are the DNA differences that distinguish species and how did these differences arise? Before this question can be addressed we must know what DNA sequences in an organism's genome are functional and how they are translated into the diversity of biological functions seen in nature. Both of these questions can now, at least to some extent, be answered through the comparison of multiple genome sequences. Although the methods of analysis have become quite sophisticated, the idea behind this approach is quite simple. Functional DNA sequences should be conserved over time and shared among closely related species, whereas non-functional or neutral sequences are free to change. This approach has been particularly useful at identifying protein coding sequences within a genome and will hopefully be as useful in identifying functional non-coding sequences. However, even with all coding and regulatory DNA sequences defined between two species, only a fraction of the DNA differences are relevant to the species' biological differences. For instance, it is well known that many changes in a protein coding region can change the amino acid sequence of a protein without affecting its function. The development of statistical methods used to infer whether changes in the amino acid sequence of a protein are functional or neutral has been of interest to both human geneticists interested in deleterious substitutions and evolutionary geneticists interested in adaptive substitutions.

The neutral theory of molecular evolution provides an essential framework in which both functional DNA sequences can be defined and functional changes can be identified. The neutral mutation random drift hypothesis was proposed independently in 1968 by Kimura [45] and in 1969 by King and Jukes [49]. The hypothesis was that the vast majority of DNA polymorphism within a species and divergence between species is neutral or non-functional with respect to fitness. Since its proposal, it has and still is intensely debated as to what are the relative contributions of positively selected and neutral mutations to DNA polymorphism and divergence. Positively selected mutations concur a fitness advantage and are rapidly fixed whereas neutral mutations follow a stochastic process of genetic drift through a population. Regardless of the actual contribution of selection and drift, the neutral theory has provided an invaluable theoretical framework in which both neutral and selective models of molecular evolution can be tested. In its simplest formulation mutations occur in a finite population of size N with rate μ per

generation. Assuming the effective population size, N, and mutation rate, μ, remain constant, at mutation drift equilibrium the rate of molecular evolution $k = \mu$, and the expected per site heterozygosity in a population under the infinite sites model, $H = 4N\mu$ [92]. Mutations which cause functional changes and are deleterious to an organism are assumed to be eliminated from a population and so do not contribute to either DNA polymorphism or divergence. As will be discussed, relaxing this latter assumption is quite important to understanding molecular evolution and is the main point of the nearly neutral theory proposed by Ohta [66]. However, it should be noted that despite known violations of even the simplest formulation of the neutral model, it adequately describes many important features of DNA polymorphism and divergence data [47].

In this review, we will examine empirical and theoretical work on how mutation, selection and drift affect the molecular evolution of protein coding sequences, and how, with the proper controls for these forces, amino acid changes with functional consequences and particularly those driven by positive selection can be identified. Although most research has been limited to protein coding DNA, much of the theory and methods which will be discussed are also applicable to non-coding DNA sequences.

2 Divergence

The protein sequence of hemoglobin and cytochrome c from multiple species enabled the first estimates of the rate of protein evolution and indicated that while each protein has its own rate of amino acid substitution the rate is constant across phylogenetic lineages [105]. Subsequent work made it clear that functionally important sites evolve more slowly than average [49] [13] and amino acids with similar physicochemical properties are substituted more often than dissimilar amino acids [105]. These observations are compatible with the neutral theory, under which functionally important amino acid positions in a protein remain constrained while neutral substitutions constitute the bulk of protein evolution. Two questions that immediately arise are: to what extent can protein sequence diverge while protein function remains the same, and to what extent do proteins' function change? The answers to these questions can, to some extent, be obtained from detailed characterization of the rate of amino acid substitution within a protein over time compared to the rate of substitution within neutral or non-functional sequences. Of par-

ticular utility is a phylogenetic approach wherein an increase in the rate of amino acid substitution in a protein along a single lineage of a phylogeny is indicative of a change in selective constraint. Although mutation rate heterogeneities have not yet been fully characterized, the genome sequences of closely related species will enable the full potential of this approach to be realized.

2.1 Theory

The rate of molecular evolution, or rate of DNA sequence divergence between species, is a function of the rate of neutral, deleterious and advantageous mutations, their selection coefficients and the effective population size. Assuming mutations act independently of one another, the expected rate of substitution is equal to the per generation influx of new mutations into a population times their probability of fixation. In a randomly mating population of constant size the probability of fixation is $(1-e^{2s})/(1-e^{-4Ns})$, where N is the effective population size and $2s$ is the selection coefficient of the homozygote [44]. Thus, the probability of fixation of a neutral substitution is $1/2N$ and the rate of neutral substitutions is μ since every generation $2N\mu$ new neutral mutations arise in a diploid population of size N. The rate of adaptive substitutions is approximately $4Ns\mu_a$, where μ_a is the mutation rate to advantageous alleles [38]. The relative rate of selected compared to neutral substitutions is shown in Figure 1. This result underlies one of the major tenets of the neutral theory: functionally important sites will remain constrained over time with high probability whereas neutral sites will evolve at a much faster rate determined by the mutation rate. In an evolutionary framework, function is defined with respect to fitness and functionally constrained sites are defined as those for which $4Ns << -1$. Thus, functional sites which when lost confer a fitness loss of as little as 0.1% are expected to be constrained even in humans who have a small effective population size, $\approx$ 13,000 [103]. These results suggest that sites under positive and negative selection can be identified by their having a rate of evolution greater than or less than, respectively, the rate of neutrally evolving sites. However, mutation rate heterogeneities, either within or between genomes, and fluctuations in effective population size also influence the rate of molecular evolution and so must be carefully accounted for. For instance, it is critical to control for mutational heterogeneities in order to distinguish between mutational coldspots and functionally constrained sites in cross genome comparisons.

2.2 ka/ks test

In protein coding sequences positive or negative selection can be tested for by a rate of amino acid substitution greater than or less than the neutral substitution rate, respectively. The ka/ks test does this by comparing the rate of amino acid substitutions to the rate of synonymous substitutions, which are assumed to be neutral. Synonymous substitutions are those that do not change the amino acid sequence of a protein and typically are found in the third but sometimes first position of a codon. Assuming synonymous sites are neutral they serve as an excellent internal control for spatial and temporal mutational heterogeneities because they are interleaved with non-synonymous, or amino acid altering sites. Although synonymous sites may not always be neutral this probably does not much affect the results of the test, as will be discussed later.

2.2.1 Estimating ka/ks

Estimating ka/ks involves two steps: estimating the effective number of synonymous and nonsynonymous (amino acid altering) sites and estimating the synonymous and nonsynonymous substitution rate from the number of synonymous and nonsynonymous differences between two sequences. These estimates require that a mutation model be specified and can be quite sensitive to the assumptions of the model. Most models assume a poisson process, i.e. mutations occur independently and with a constant rate. The simplest model is to assume equal base frequencies and equal probability of mutation among the four bases. Under this model the effective number of synonymous sites can be approximated as 1/3 the number of two-fold plus all the four-fold degenerate sites and the rate of synonymous and nonsynonymous substitution between two sequences can be estimated by $d = -\frac{3}{4}ln(1-\frac{4}{3}p)$ [43], where p is the proportion of synonymous and nonsynonymous differences, respectively. When divergence is high d is much greater than p to correct for multiple mutations at the same site and d becomes biased as sites becomes saturated with many substitutions per site. A maximum likelihood estimator of ka/ks has also been developed from the probability of substitution between codons $P(t) = e^{Qt}$ where Q is the rate matrix and t is time [32]. The maximum likelihood estimate is often quite similar to the approximate estimate for low levels of divergence but is not biased when divergence is high [100].

Estimates of ka/ks are quite sensitive to the underlying mutation model

that is assumed. Most importantly, differences in the rate of transition and transversion mutations must be accounted for since synonymous mutations are more often transitions than transversions. Unequal base composition is often found at first, second and third positions in a codon, reflecting both mutational and selective forces [55]. For amino acid altering sites selection is likely the dominant force, whereas for synonymous sites both forces have been shown to be influential. Comparison of base composition at synonymous and intergenic sites, assumed to be neutral, shows codon usage bias, or biased base composition among synonymous codons, can in part be explained by mutational biases which are expected to affect both coding and noncoding sequences [40] [88]. However, codon bias is often more extreme than base composition biases in surrounding regions and this remaining codon bias is best explained by weak selection $|2Ns| \approx 1 - 3$ acting on translational accuracy or efficiency or some other character affecting fitness such as mRNA secondary structure [1].

Both maximum likelihood and approximate methods have been developed to account for mutational biases [100]. These range from a two parameter model, which accounts for different rates of transitions and transversions, to a 61 parameter model which accounts for unequal usage of all codons. Substantial biases in ka/ks estimates are obtained when the incorrect mutational model is used [100]. A common observation is a high G+C content at third positions within a codon. It is easy to see that as third positions reach saturation the number of differences per site is expected to be greater than the maximum of 3/4 expected under the Jukes-Cantor model [43], and produce an overestimate of the substitution rate. Because mutation parameters are typically not known they must be estimated from the data. These are then used to calculate both the effective number of synonymous and nonsynonymous sites and estimate the number of synonymous and nonsynonymous substitutions. When there are multiple substitutions within a codon more than one order of events is possible and the number of synonymous and nonsynonymous substitutions depend on this order. Most methods estimate the probability of a nonsynonymous compared to synonymous substitution from codons with only a single change and then weight the order of events by their probability of occurrence [100]. For this reason and because of multiple mutations at a single site, ka/ks estimates are not reliable when either ka or ks is greater than one.

More complex mutation models have also been developed to account for variable mutation rates across sites [31]. Accounting for CpG sites, which

mutate at a rate 10 to 15 times higher than non-CpG sites in humans [37] [50], is important to estimating substitution rates [70] but is not commonly done. Finally, nonstationary nucleotide content can bias ka/ks estimates but is rarely incorporated into substitution models (but see [29]).

2.2.2 Application of the ka/ks test

The ka/ks test has been applied to numerous genes and is quite useful to understanding the selective constraints acting on the encoded proteins as well as any changes in selective constraints. A ka/ks of 0.20 can be interpreted as 80% of amino acid altering mutations within the protein being deleterious or put differently 80% of the amino acid positions being functionally constrained. The average ka/ks between human and rodent is 0.18 using 1880 orthologous genes [58], 0.15 using 2112 genes [41], and the median of 12,615 human and mouse orthologues is 0.12 [89]. The average ka/ks along the lineage leading to *D. melanogaster* is 0.20 and to *D. simulans* is 0.12 from 44 genes [18]. Between *Escherichia coli* and other bacteria the average ka/ks is 0.08 from 3106 genes [42]. Thus, most proteins are tightly constrained. The magnitude of the difference between the synonymous and nonsynonymous substitution rates has been shown to be quite useful for identifying exons from divergence data [62].

From genomic studies and from surveys of the literature [16] [56] less than 1% of genes have a ka/ks ratio that is significantly higher than one. Most of these genes are involved in sexual selection or disease resistance [98]. These results indicate that most proteins are highly constrained and only a few proteins evolve rapidly under positive selection for a change in protein function. However, the ka/ks test, while robust, is likely too conservative in detecting proteins that have evolved under positive selection between two species. The reason is that some regions of constraint within a protein are likely maintained during the evolution of a new or improved function of a protein. These regions will lower the overall rate of amino acid substitution within a protein below the neutral rate unless the adaptive regions evolve at a rate fast enough to bring the average ka/ks of the entire protein above one. Alternatively, the period of adaptive evolution and rapid amino acid substitution may only occur for a short time period followed by selective constraint on the new or improved protein. For instance, the *Odysseus* gene, identified by its involvement in reproductive isolation in *Drosophila*, has undergone 7 amino acid changes in its homeodomain in 700 million years of

divergence, whereas 10 amino acid substitutions have occurred in a half millions years along the lineage leading to *D. mauritiana* [86]. Estimation of negative selection is also inhibited if spatial and temporal constraints are not accounted for. A number of methods have been developed to account for positive or negative selection limited in time or to a subset of amino acid positions within a protein.

2.2.3 Codon based ka/ks tests

An alternative to estimating the average ka/ks of an entire protein, ka/ks can be estimated for protein domains or codons within a protein. Doing so has greatly facilitated both identifying proteins under positive selection and better describing functional constraints on a protein [5] [64] [102] [98]. Both a maximum likelihood and maximum parsimony method have been implemented to account for heterogeneous selection pressure among sites.

Within a maximum likelihood framework codons can evolve at different rates under a variety of mutation models [32]. In general it is assumed that a fraction of amino acid positions within a protein are constrained $ka/ks < 1$, a fraction are neutral $ka/ks = 1$, and a fraction are under positive selection $ka/ks >1$. Positive selection is detected if the likelihood ratio test indicates there is a significant improvement in the fit of the model to the data when the fraction of sites evolving under positive selection is greater than zero. Identification of the sites within each fraction is possible using a bayesian approach [64]. A different approach is to use a maximum parsimony phylogeny to infer ancestral states of a sequence and then estimate ka/ks for each codon within a protein [78][24].

In general, both maximum likelihood and maximum parsimony methods provide vast improvements in describing the effects of both positive and negative selection on a protein, however, there are some drawbacks. The maximum likelihood method may depend on the initialization of the algorithm due to multiple local maxima in the likelihood surface [79] and may also have a high rate of false positives due to assumptions of the methodologies [80]. The maximum parsimony method does not incorporate many mutational biases and codon usage bias. While both approaches appear to have reasonable power and reliability [3] [4] [79] [80], a large number of sequences that are not too close or distantly related are required for such analyses.

Codons can also be categorized by the physical and chemical properties of the amino acids they encode, and amino acid substitutions can be classi-

fied as conservative and radical based on these properties [33]. Across many proteins, the rate of radical amino acid substitutions is much slower than that of conservative amino acid substitutions reflecting the greater strength of purifying selection on changes that are more likely to affect the structure of a protein [105] [34] [33] [95]. This classification also assists in the detection of positive selection. Comparison of human and old world monkey orthologues related to male reproduction revealed a significantly higher rate of conservative amino acid substitutions to synonymous substitutions among the 11 most rapidly evolving proteins. In contrast, the most rapidly evolving proteins unrelated to male reproduction showed constraint on both conservative and radical substitutions [95].

2.2.4 Lineage specific ka/ks test

If a protein has experienced positive selection for a new or modified function, the rate of protein evolution in the lineage subject to positive selection is expected to be higher than other phylogenetic lineages. By constructing a phylogeny and inferring the ancestral states of a protein, the ka/ks ratio can be tested along each lineage of a phylogeny [61] [104] [25]. The drawback of this approach is that it requires a true phylogeny and except for the case of three species most phylogenies have considerable uncertainty. To avoid this uncertainty a maximum likelihood approach was developed to estimate ka/ks for each branch across all probable phylogenies weighted by their likelihood [99]. Applications of lineage specific tests for positive selection show they greatly facilitate its detection [61] [97] and the combination of codon and lineage specific estimation of ka/ks [101] provides a powerful means of describing the effects of both positive and negative selection on protein evolution.

2.3 Change in ka/ks

If the effective population size and the selective constraint on a protein remains constant over time, ka/ks is also expected to remain constant. An increase in ka/ks can result from a loss of constraint due to either a decrease in effective population size or a decrease in selection intensity. Positive selection can also increase ka/ks. Distinguishing between these possibilities is important to understanding what evolutionary forces drive protein evolution and why protein sequences differ in different species. A change in

effective population size can be distinguished from a change in selection intensity since a change in population size is expected to affect all genes along a lineage whereas a change in selection intensity can be different for different genes [20]. Thus, a change in the average ka/ks of many genes along a lineage can be attributed to a change in effective population size and a greater than expected variance in ka/ks can be attributed to changes in selection, either positive or negative. Whether or not most proteins vary in their rate of evolution is relevant to the molecular clock hypothesis [53], which states that nearly all proteins have a constant but protein specific rate of amino acid substitution [105].

2.3.1 Relative rates test

A change in the rate of evolution as measured by ka/ks can be tested for using the relative rates test which compares pairwise rates of evolution using three or more taxa [93]. For three taxa a, b and c where a and b are the most closely related, the distance between a and c is expected to be the same as the distance between b and c. Maximum likelihood methods can also be used to test for differences in ka/ks across lineages with the advantage that pairwise estimates of ka/ks are not needed to calculate the ka/ks ratio along different branches [99]. Using artiodactyl as an outgroup the average ka/ks along the primate lineage, 0.27, is significantly greater than along the rodent lineage, 0.17 [68]. This difference can be attributed to the presumably smaller effective population size along the primate compared to rodent lineages. In addition to a difference in the average ka/ks, the index of dispersion (variance/mean) of the substitution rate is high for both ka (5.6) and ks (5.9) [68]. For a poisson process the index of dispersion is expected to be one, but if selection has made slight but individually insignificant lineage specific changes in the rate of ka or ks the index of dispersion is expected to be greater than one [30]. In *Drosophila* the index of dispersion of ka and ks is greater than one in some but not all genes which may reflect mutational heterogeneities, codon bias, or selection [83].

The genome sequence of closely related species has made it possible to apply the relative rates test to all orthologous proteins among three genomes. Of 2112 human-mouse-rat orthologues less than 1% show a significantly different rate of evolution along one lineage [58][41]. Similar results were obtained from the comparison of closely related bacterial and archaeal species [41]. The small number of orthologous genes which have a variable rate of

evolution supports the molecular clock hypothesis and makes many proteins quite useful for phylogenetic studies.

2.3.2 Duplicate genes

In contrast to the often constant rate of evolution of orthologous genes, an increase in the rate of evolution is often observed in paralogues following gene duplication. This is expected if adaptive evolution often proceeds through gene duplication followed by the evolution of new function [65]. The comparison of orthologous and paralogous genes provides a powerful approach to inferring functional changes within a protein [35] [6]. Orthologues performing the same function should be under the same selective constraints and should have the same rate of evolution. If a duplicate has maintained the same function, the location of its conserved domains and its rate of evolution are expected to be equal to that of its orthologue and paralogue. Duplicates of this nature may arise when the expression of a protein is needed in a new tissue or stage of development without any accompanying change in the amino acid sequence. If a duplicate has evolved a new function, its rate of evolution is expected to be greater than that of its orthologue and paralogue. However, an alternative explanation for rapid evolution following gene duplication is loss of constraint due to complete or partial loss of function in one or both duplicates. If the duplicates subfunctionalize the original protein's functions, both duplicates are expected to have a higher rate of evolution [57] [27]. Distinguishing loss of constraint from rapid evolution driven by positive selection is quite difficult since loss of constraint often precedes the evolution of new function. The inference of a change in function or constraint is facilitated by examination of a protein's structure and the types of amino acid changes that distinguish paralogues [6].

Comparison of orthologues and paralogues from two pairs of closely related bacterial and three closely related eukaryotic species revealed the ka/ks of paralogues is two to three times greater than that of their unduplicated orthologue [52]. In support of the subfunctionalization model, of 105 pairs of genes, only five evolved at significantly different rates following duplication.

2.4 Genome comparisons

The genome sequences of closely related species makes it possible to quantify the frequency of positive and negative selection in the genome and address

a number of new questions. Distantly related organisms cannot easily be compared since synonymous and other unconstrained sites become saturated and only proteins under considerable constraint can be identified as orthologues. Even closely related species may contain genes that have evolved so rapidly that they are no longer easily identified [87]. The sequence of the human and mouse genome allows the neutral theory to be tested using the ka/ks test. The median ka/ks of 12,845 putative orthologues is 0.115, and few genes show ka/ks greater than one [89]. While this clearly demonstrates the strong role of purifying selection in protein evolution it does not rule out a significant contribution of adaptive substitutions due to the conservative nature of the ka/ks test.

The human and mouse genomes also make it possible to characterize the mutational biases and heterogeneities that have affected the divergence of the human and mouse genome. These can then be used to further refine mutation models and better estimate rates of neutral divergence either at synonymous, nonsynonymous or non-coding sites. Not only is this important to estimating ka/ks but also to identifying conserved non-coding sequences.

The most important parameter to estimate before function can be inferred from sequence constraint or divergence is the mean and variance in the rate of neutral substutitions. For a neutral sequence, the divergence between two species is the sum of the divergence that occurred since the split of the two species, t, and the divergence between the two alleles in the ancestral population that went on to become fixed in the two species (Figure 2). The expectation of these two quantities is $2\mu t + 4Nu$, where t is time in generations, $2\mu t$ is the divergence that occurred after speciation and $4N\mu$ is the divergence due to segregation of ancestral polymorphism before speciation. While the variance of $2\mu t$ is assumed to be poisson, the variance of $4N\mu$ is greater than poisson since it includes the evolutionary variance inherent to the coalescence process [82]. When there is recombination, each locus in the genome is expected to have its own genealogical history and so $4N\mu$ is expected to be different for different genes in the genome. Furthermore, because reproductive isolation is not instantaneous, different regions of the genome may become incompatible between species before others [85]. Thus, even in the absence of selective constraint and heterogeneous mutation rates across the genome, rates of divergence can be quite variable across the genome. This makes it difficult to infer selective constraint for a non-coding region with low levels of divergence.

The comparison of the human and mouse genome show substantial vari-

ability in rates of divergence across the genome [89]. In human baboon comparisons there is more variation than expected from a poisson process at both small, 10bp, and large, 100kb, scales [74]. Two approaches were taken to estimate the mean and variance of the neutral substitution rate between human and mouse. The first was to estimate the distribution of ks from coding sequences and the second was to estimate ks from repetitive elements as compared to their ancestral consensus sequence. This second approach, however, assumes that G+C content is stationary over time. If a transposable element with low G+C content arrives in a region of high G+C content, evolution to high G+C content may be rapid and may lead to biased estimates of the substitution rate [36]. So long as variation in the neutral substitution rate occurs on a scale larger than that of the length of functional conserved non-coding elements, hidden markov models should be able to predict conserved regions using the surrounding region to estimate the neutral substitution rate.

2.5 Independence

Thus far it has been assumed that amino acid changes within a protein occur independently of one another. The covarion model supposes that amino acid substitutions are not independent of one another [26]. The degree of dependence, or epistasis among amino acid substitutions, can be defined as the average fraction of all codons within a protein whose state determines the fitness effect of an amino acid substitution. Models which assume an amino acid substitution can affect the fitness effects of subsequent substitutions have been studied [81] and have been found to increase the variance in the rate of evolution [69]. The fact that many human pathogenic amino acid substitutions are present in non human species suggests numerous epistatic interactions either within or between proteins [51]. Comparison of the human and mouse genome revealed 160 examples of such from 7,293 disease associated amino acid mutations [89]. From 32 proteins with numerous pathogenic alleles defined it was estimated that approximately 10% of amino substitutions occur at sites known to cause pathogenesis [51]. This estimate is independent of sequence divergence and has broad implications for studies of molecular evolution which typically assume no fitness interactions among sites (but see [40]). Epistasis is also known to occur between proteins, in which case substitutions in one protein influence the fitness effects of a substitution in a second protein. These types of interactions are thought to play

an important role in the Dobzhansky-Muller hybrid incompatibilities that distinguish species [94]. To take advantage of these types of interactions a method of detecting positive selection was developed which uses character states and the subsequent number of amino acid changes along a lineage to infer positive selection [12].

2.6 Selection on synonymous sites

The inference of positive or negative selection based on the ka/ks test assumes changes at synonymous sites are neutral. In bacteria, yeast, nematode and flies, codon usage bias ranges from highly biased genes to genes with almost no bias and this bias is presumably caused by translational accuracy and/or efficiency since it is correlated with levels of gene expression [2]. In primates and rodents there is no support for selection acting on synonymous sites, an observation that can be explained by their smaller effective population size [88]. Using population genetic theory and *Drosophila* data, the frequency of preferentially used codons can be explained by an intensity of selection,/$2Ns$, between 0.1 and 3 [1]. The critical issue is the extent to which the synonymous substitution rate is increased or decreased due to positive or negative selection on codon usage bias. For weak selection, $2Ns \approx 1$, selection on synonymous sites will have very little effect on the rate of substitution (Figure 1). Previous studies using approximate estimators of ks found a strong negative correlation between codon bias and ks, however, maximum likelihood estimators of ks show no correlation between ks and codon bias in *Drosophila* [15] [60]. Thus, so long as ks remains unaffected by selection on synonymous sites it may be used as a good approximation of the neutral substitution rate.

3 Polymorphism

The ability to survey protein polymorphism provided the opportunity of associating protein polymorphism with phenotypic variation and/or natural selection [54]. Despite surveys in numerous species most protein polymorphism appeared neutral with respect to fitness and function [47]. The only evidence for abundant functional polymorphism under selection came from amino acid mutations found at very low frequencies in a population, $< 1\%$, and were interpreted as deleterious mutations kept at low frequencies by pu-

rifying selection [67]. However, the inference of deleterious mutations was confounded by changes in population size which could also explain the data [67] [90]. By examining the difference in the frequency spectrum of nonsynonymous to synonymous mutations within a population the effects of a population's demographic history are removed, as both nonsynonymous and synonymous variation are affected by demographics. Thus far, both humans [21], *D. melanogaster* [22] and *E. coli* [73] contain a large amount of low frequency amino acid polymorphism which cannot be explained by demographics. Of particular relevance to humans is the frequency distribution and identity of mutations (i.e. coding or non-coding) that contribute to human genetic diseases and phenotypes. If a large fraction of these mutations are deleterious with respect to fitness they are expected to reside at low, $< 10\%$, frequencies in the population (see below). In contrast, if these mutations are neutral they are expected to reside at much higher frequencies on average, as proposed by the common disease common variant hypothesis [72]. The relationship between fitness and human health can now be examined by comparing amino acid substitutions inferred to be deleterious from polymorphism or divergence data to those found to be association with human genetic diseases. While it is possible to make inferences on a collection of amino acid polymorphism, the identification of particular amino acid polymorphism under negative [75] and especially positive selection [19] is greatly facilitated using linked neutral polymorphism, which is also expected to be affected by selection.

3.1 Theory

Levels of DNA polymorphism within a population are a complex function of mutation rate, effective population size, population history and selection. For a population of constant size under the infinite sites model, the expected per site heterozygosity or proportion of differences between two sequences is $4N\mu$, and the frequency spectrum of segregating sites is given by $\phi(x)dx = \frac{4N\mu}{x}dx$ [92] [46]. The frequency spectrum of segregating sites reflects the balance between an average of $2N\mu$ mutations which enter the population at a frequency of $1/2N$ every generation and drift which results in the loss or fixation of mutations. Positive selection increases rates of polymorphism and results in more high frequency mutations compared to neutral mutations. Negative selection removes mutations from a population and results only in low frequency polymorphism. Thus, both positive and negative selection produce a

skew in the frequency spectrum in comparison to neutral polymorphism (Figure 3). The accumulation of non-neutral sites at low and high frequencies reflects the greater efficacy of selection at intermediate frequencies compared to low or high frequencies where the sampling variance is high and substantial drift of non-neutral mutations can occur. The influence of selection becomes stronger than drift when the frequency of a mutation is greater than $1/4Ns$ [17]. Thus, slightly deleterious mutations can reach higher frequencies than strongly deleterious mutations, and in contrast to divergence data, selective constraint can be quantified as a function of selection intensity.

3.2 Detecting selection

Positive and negative selection can be detected from the ratio of amino acid to synonymous polymorphism, conceptually equivalent to the ka/ks test. However, the ratio of amino acid to synonymous polymorphism changes as a function of frequency and as a function of the intensity of selection. Considering just the effects of purifying selection, the ratio of amino acid to synonymous polymorphism at a frequency of $1/2N$ is expected to be close to one if the rate of polymorphism is measured using the effective number of nonsynonymous and synonymous sites. Dominant lethals are eliminate in a single generation. A longer period of time is needed to eliminate deleterious mutations and as a consequence they attain higher frequencies. Neutral mutations are only eliminated by chance. Thus, the ratio of amino acid to synonymous polymorphism is expected to be close to one at very low frequencies and gradually decreases until only neutral polymorphism remains, at which point the ratio of the per site rate of amino acid to synonymous variation is expected to be equal to the selective constraint on a protein. Of course when positive selection is present the ratio of amino acid to synonymous variation is expected to increase from intermediate to high frequencies as well as for divergence (see below).

The frequency distribution of amino acid polymorphism can be used to estimate the strength of positive or negative selection [73] and the fraction of polymorphism which is neutral [67] [47]. Strongly deleterious mutations, such as those causing severe human genetic diseases, are kept at very low frequencies in a population whereas slightly deleterious mutations, such as those contributing to complex human genetic diseases, are able to drift to higher and even detectable frequencies ($> 1\%$) in a population. Strongly advantageous mutations tend to lie at very low or high frequencies (Figure 3).

Because amino acid polymorphism under selection tends to lie at either low or high frequencies the ratio of amino acid to synonymous polymorphism at intermediate frequencies is a slight overestimate of the fraction of nonsynonymous mutations which are effectively neutral, assuming no overdominance. The difference in the ratio of amino acid to synonymous polymorphism at intermediate compared to low or high frequencies can be attributed to selection [21].

3.3 Application to data

Polymorphism data is particularly useful for understanding deleterious mutations because advantageous mutations spread quickly through a population and so are rare compared to neutral mutations, and because deleterious mutations are expected to be common at low frequencies in a population. Since the earliest allozyme studies a substantial excess of low frequency amino acid polymorphism was noted and attributed to slightly deleterious mutations [67]. However, demographic effects such as an increase in population size could also explain this excess of low frequency variation. A significantly higher ratio of amino acid to synonymous variation at low compared to intermediate frequencies cannot be explained by an increase in population size since both nonsynonymous and synonymous polymorphic sites should be similarly affected. From two different polymorphism surveys totaling 181 genes, it was estimated that a half of low frequency (1-10%) amino acid polymorphism is deleterious (a third of all amino acid altering single nucleotide polymorphism, SNPs) and the average number of deleterious amino acid mutations carried by an individual was estimated to be at least 500 [21]. Similar estimates were obtained by comparing population specific and shared SNPs with the logic that neutral but not deleterious SNPs are able to migrate across populations. The large fraction of slightly deleterious amino acid polymorphism has implications for our understanding of rates and patterns of molecular evolution since it is these mutations which first become effectively neutral and are able to fix in a population with a smaller effective size. These results are also relevant to the assumption that many common complex human genetic diseases are caused by common alleles in a population [72] [28] whereas both theoretical and empirical considerations suggest nearly all of these alleles likely reside at a frequency of less than 10% [21] [71].

A number of other methods have also been devised to determine what

fraction of amino acid polymorphism is slightly deleterious. One approach is to determine whether an amino acid polymorphism has functional consequences based on protein structure annotations such as the location of active or binding sites and disulfide bonds, or based on physical and chemical properties of an amino acid substitution such as the hydrophobicity and electrostatic charge change and the effect on protein solubility. These measures must be calibrated using amino acid changes known to affect function. Calibration can be obtained from amino acid changes annotated as causing a human genetic disease [77] or from studies of the *lac* repressor or lysozyme proteins for which function has been measured for nearly all possible amino acid substitutions [11]. Using both structural and divergence data, estimates of the fraction of amino acid polymorphism that is deleterious ranges from 20% [77] to 29% [11] and an individual is expected to carry 10^3 to 10^4 of these mutations in their genome, respectively.

Another approach relies entirely on divergence data with the logic that amino acid sites conserved over time are likely functional and deleterious when mutated [63]. However, the opposite conclusions were reached using this method; very few amino acid SNPs were found to be damaging to the extent of affecting human health since 20% of SNPs were predicted to affect function and the estimated rate of false positives was also 20% [63]. The higher estimates of previous studies were attributed to not accounting for false positives [11] and to estimates based on SNPs biased to an unrepresentative set of genes [77]. However, the 20% estimate of the rate of false positives comes from a single protein, the *lac* repressor, and it is not clear that this estimate is applicable to other proteins.

The different estimates of functional or deleterious amino acid polymorphism likely stem from an important point relating to the definitions of "deleterious", "functional", and "human health", which can be defined as follows. Deleterious mutations are those that affect fitness, are removed from a population and rarely contribute to protein divergence. Functional mutations are those that in the lab produce a detectable phenotype. Mutations affecting human health are those that contribute to human genetic diseases. While these categories of mutations obviously overlap their relationships are not easily defined. The fraction of amino acid SNPs that affect human health should be directly estimated. A further complication of using divergence data to infer function is that many human disease alleles have been shown to be present in mouse [51]. While constraint estimated from polymorphism data does not suffer from this limitation there is no clear relationship between the

fitness consequence of a mutation and its affect on human health, although clearly the two must be correlated.

4 Polymorphism and Divergence

The comparison of polymorphism to divergence data presents the most powerful means of disentangling the selective and demographic forces governing protein evolution. The comparison is also the most difficult. One of the first and now most widely used comparisons of polymorphism and divergence is the McDonald-Kreitman test [59]. Originally proposed as a test for positive selection based on an excess of amino acid divergence compared to that expected based on levels of polymorphism, the test is equally capable of detecting negative selection based on an excess of amino acid polymorphism compared to divergence [84]. In fact, a higher ratio of polymorphism compared to divergence is observed in a number of mitochondrial genomes and is interpreted as segregating deleterious amino acid polymorphism [91]. Positive and negative selection can be distinguished using frequency to infer the contribution of negative selection to polymorphism [22]. As previously discussed (ıSection 2.3), changes in effective population size and selective constraint are expected to first change the ratio of amino acid to synonymous polymorphism and subsequently the ratio of amino acid to synonymous divergence. Distinguishing between the influences of positive negative selection and drift will provide the most meaningful understanding of how mutation, selection and drift interact within natural populations and give rise to genome differences. The recent expansion in human population size in combination with changes in selective constraint provides a unique opportunity to address these issues in humans.

4.1 Detecting selection

The McDonald-Kreitman test is a test of independence between the number of nonsynonymous and synonymous polymorphic sites to the number of nonsynonymous and synonymous fixed differences between species [59]. If all mutations are neutral the ratio of nonsynonymous to synonymous polymorphism is expected to be equal that of divergence. Positive selection is expected to increase the number of amino acid substitutions but have little impact on polymorphism (Figure 1). Negative selection is expected to af-

fect amino acid polymorphism but not divergence. A change in population size is expected to have a dynamic effect first on polymorphism and then on divergence. As this is a non-stationary process no theoretical models of the process have been developed. However, the magnitude of the effect is expected to be a function of the fraction of amino acid polymorphism with fitness effects that becomes effectively neutral for a given change in effective population size. The larger a decrease in effective population size, the larger the fraction of amino acid mutations which become effectively neutral and the larger the increase in ka. The large fraction of amino acid polymorphism found at frequencies of 1-10% in both *D. melanogaster* [22] and humans [21] suggests that the effects of a change in population size on the rate of amino acid substitution may be quite large. This makes it difficult to distinguish between positive selection and a historical change in selective constraint which can both produce a ratio of amino acid to synonymous divergence higher than that of polymorphism [22].

To distinguish positive and negative selection from changes in population size, polymorphism at different frequencies and divergence must be compared at multiple loci [22]. Because the bulk of intermediate frequency amino acid polymorphism is likely neutral it can be used to gauge the contribution of deleterious mutations to polymorphism by the ratio of amino acid to synonymous variation at low compared to intermediate frequencies. Positive selection can be distinguished from a change in population size by examining numerous genes since all genes should be affected by a change in population size but only a small fraction of genes are likely under positive selection [22].

A maximum likelihood method of estimating the strength of positive or negative selection has been developed based on the frequency spectrum expected in an equilibrium population and divergence between species [9]. However, the method must assume an equilibrium population and the data is fit to only a single selection coefficient, so positive and negative selection are confounded. This framework has now been extended to estimate the distribution of either positive or negative selection coefficients [7].

4.2 Application to data

A number of species now have polymorphism and divergence data at multiple loci. The maximum likelihood estimates of $4Ns$ from 12 *Arabidopsis* genes is between -2 and 1 and from 32 *Drosophila* genes is between -1 and 4 [8]. An approximate fit of the excess of amino acid polymorphism found in

humans to that expected in an equilibrium population produced estimates of $4Ns$ between -10 and -1000 [21]. The excess of amino acid divergence between *D. melanogaster* and *D. simulans* compared to that expected based on polymorphism suggested that about 1/3 of amino acid substitutions were driven by positive selection [22] [76]. While a change in population size can not account for the entire excess of amino acid divergence it may have had some impact [22]. Sampling of genes in other species will determine the contribution of adaptive substitutions to divergence since a numerous species are not likely to have the same demographic history.

5 Conclusions

The number of questions limited by lack of divergence data is rapidly growing smaller. The abundance of divergence data has lead to more accurate mutation models which are essential for estimating functional and fitness consequences of amino acid mutations. However, mutation rate parameters have yet to be fully characterized with respect to their variation within a genome [74] [50] and between genomes [96]. Despite these uncertainties, we have refined our methods of inference to the point where sites inferred to be under positive selection or sites constrained in some orthologues or paralogues but not in others should be experimentally tested.

Polymorphism data is also now available on a genomic level, although limited in form. Large samples from multiple loci are needed to control for demographic effects and better understand how purifying selection translates into functional constraint. The intense focus on functional human polymorphism will no doubt put to use and require improvements on methods of inferring selection on amino acid polymorphism. However, the best understanding of a protein's evolution no doubt comes from the analysis of both polymorphism and divergence data.

With nearly 40 years of protein evolution studies, the next frontier lies in the study of non-coding sequences and their regulatory functions. Although most regulatory sequences have not been identified, constraint in non-coding regions between two genomes provides a fast method of identifying candidate regulatory sequences once the genomes of closely related species are made available.

Non-coding regions can by analyzed using the same types of methods applied to coding regions. From human polymorphism surveys rates of poly-

morphism in 5' UTR, intron and 3' UTR regions were found to be half the rates found at synonymous sites [39] [10]. In contrast, rates of divergence at 5' UTR, 3' UTR and synonymous sites were similar, suggesting no or little selective constraint [58]. Many explanations are plausible but it should be noted that different approximate methods were used to estimate rates of variation. Examination of rates of evolution between human and mouse in known regulatory sequences revealed substitution rates in transcription factor binding sites are 2/3 the rate of background sequences [14]. However, the substitutions found in binding sites resulted in more than one third of the sites being disrupted in one of the two species suggesting transcription factor binding sites may have a high rate of turnover. Further work will no doubt clarify the strength of selective forces acting on regulatory elements and their contribution to human genetic diseases and adaptive evolution.

References

[1] Akashi H. 1995. Inferring weak selection from patterns of polymorphism and divergence at "silent" sites in Drosophila DNA. *Genetics* 139:1067–1076.

[2] Akashi H. 2001. Gene expression and molecular evolution. *Curr. Opin. Genet. Dev.* 11:660–666.

[3] Anisimova M, Bielawski J, Yang Z. 2001. Accuracy and power of the likelihood ratio test in detecting adaptive molecular evolution. *Mol. Biol. Evol.* 18:1585–1592.

[4] Anisimova M, Bielawski J, Yang Z. 2002. Accuracy and power of bayes prediction of amino acid sites under positive selection. *Mol. Biol. Evol.* 19:950–958.

[5] Bielawski J, Yang Z. 2001. Positive and negative selection in the DAZ gene family. *Mol. Biol. Evol.* 18:523–529.

[6] Blouin C, Boucher Y, Roger A. 2003. Inferring functional constraints and divergence in protein families using 3D mapping of phylogenetic information. *Nuclei Acids Res.* 31:790–797.

[7] Bustamante C, Nielsen R, Hartl D. 2003. Maximum likelihood and bayesian method for estimating the distribution of selective effects among classes of mutations using DNA polymorphism data. *Theor. Popul. Biol.* in press.

[8] Bustamante C, Nielsen R, Sawyer S, Olsen K, Purugganan M, Hartl D. 2002. The cost of inbreeding in Arabidopsis. *Nature* 416:531–534.

[9] Bustamante C, Wakeley J, Sawyer S, Hartl D. 2001. Directional selection and the site-frequency spectrum. *Genetics* 159:1779–1788.

[10] Cargill M, Altshuler D, Ireland J, Sklar P, Ardlie K, Patil N, Shaw N, Lane C, Lim E, Kalyanaraman N, Nemesh J, Ziaugra L, Friedland L, Rolfe A, Warrington J, Lipshutz R, Daley G, Lander E. 1999. Characterization of single-nucleotide polymorphisms in coding regions of human genes. *Nat. Genet.* 22:231–238.

[11] Chasman D, Adams R. 2001. Predicting the functional consequences of non-synonymous single nucleotide polymorphisms: structure-based assessment of amino acid variation. *J. Mol. Biol.* 307:683–706.

[12] Creevey C, McInerney J. 2002. An algorithm for detecting directional and non-directional positive selection, neutrality and negative selection in protein coding DNA sequences. *Gene* 300:43–51.

[13] Dayhoff M. 1972. *Atlas of Protein Sequence and Structure*, volume 5. Washington, DC: National Biomedical Research Foundation.

[14] Dermitzakis E, Clark A. 2002. Evolution of transcription factor binding sites in Mammalian gene regulatory regions: conservation and turnover. *Mol. Biol. Evol.* 19:1114–1121.

[15] Dunn K, Bielawski J, Yang Z. 2001. Substitution rates in Drosophila nuclear genes: implications for translational selection. *Genetics* 157:295–305.

[16] Endo T, Ikeo K, Gojobori T. 1996. Large-scale search for genes on which positive selection may operate. *Mol. Biol. Evol.* 13:685–690.

[17] Ewens W. 1979. *Mathematical Population Genetics.* Springer-Verlag.

[18] Eyre-Walker A, Keightley P, Smith N, Gaffney D. 2002. Quantifying the slightly deleterious mutation model of molecular evolution. *Mol. Biol. Evol.* 19:2142–2149.

[19] Fay J, Wu C. 2000. Hitchhiking under positive Darwinian selection. *Genetics* 155:1405–1413.

[20] Fay J, Wu C. 2001. The neutral theory in the genomic era. *Curr. Opin. Genet. Dev.* 11:642–646.

[21] Fay J, Wyckoff G, Wu C. 2001. Positive and negative selection on the human genome. *Genetics* 158:1227–1234.

[22] Fay J, Wyckoff G, Wu C. 2002. Testing the neutral theory of molecular evolution with genomic data from Drosophila. *Nature* 415:1024–1026.

[23] Fisher R. 1930. *The Genetical Theory of Natural Selection.* Oxford: Clarendon Press.

[24] Fitch W, Bush R, Bender C, Cox N. 1997. Long term trends in the evolution of H(3) HA1 human influenza type A. *Proc. Natl. Acad. Sci. USA* 94:7712–7718.

[25] Fitch W, Leiter J, Li X, Palese P. 1991. Positive Darwinian evolution in human influenza A viruses. *Proc. Natl. Acad. Sci. USA* 88:4270–4274.

[26] Fitch W, Markowitz E. 1970. An improved method for determining codon variability in a gene and its application to the rate of fixation of mutations in evolution. *Biochem. Genet.* 4:579–593.

[27] Force A, Lynch M, Pickett F, Amores A, Yan Y, Postlethwait J. 1999. Preservation of duplicate genes by complementary, degenerative mutations. *Genetics* 151:1531–1545.

[28] Gabriel S, Schaffner S, Nguyen H, Moore J, Roy J, Blumenstiel B, Higgins J, DeFelice M, Lochner A, Faggart M, Liu-Cordero S, Rotimi C, Adeyemo A, Cooper R, Ward R, Lander E, Daly M, Altshuler D. 2002. The structure of haplotype blocks in the human genome. *Science* 296:2225–2229.

[29] Galtier N, Tourasse N, Gouy M. 1999. A nonhyperthermophilic common ancestor to extant life forms. *Science* 283:220–221.

[30] Gillespie J. 1989. Lineage effects and the index of dispersion of molecular evolution. *Mol. Biol. Evol.* 6:636–647.

[31] Goldman N, Whelan S. 2000. Statistical tests of gamma-distributed rate heterogeneity in models of sequence evolution in phylogenetics. *Mol. Biol. Evol.* 17:975–978.

[32] Goldman N, Yang Z. 1994. A codon-based model of nucleotide substitution for protein-coding DNA sequences. *Mol. Biol. Evol.* 11:725–736.

[33] Grantham R. 1974. Amino acid difference formula to help explain protein evolution. *Science* 185:862–864.

[34] Graur D. 1985. Amino acid composition and the evolutionary rates of protein-coding genes. *J. Mol. Evol.* 22:53–62.

[35] Gu X. 2001. Maximum-likelihood approach for gene family evolution under functional divergence. *Mol. Biol. Evol.* 18:453–464.

[36] Gu Z, Wang H, Nekrutenko A, Li W. 2000. Densities, length proportions, and other distributional features of repetitive sequences in the human genome estimated from 430 megabases of genomic sequence. *Gene* 259:81–88.

[37] Hacia J, Fan J, Ryder O, Jin L, Edgemon K, Ghandour G, Mayer R, Sun B, Hsie L, Robbins C, Brody L, Wang D, Lander E, Lipshutz R, Fodor S, Collins F. 1999. Determination of ancestral alleles for human single-nucleotide polymorphisms using high-density oligonucleotide arrays. *Nat. Genet.* 22:164–167.

[38] Haldane J. 1927. The mathematical theory of natural and artificial selection. Part V. *Proc. Cambridge Philos. Soc.* 23:838–844.

[39] Halushka M, Fan J, Bentley K, Hsie L, Shen N, Weder A, Cooper R, Lipshutz R, Chakravarti A. 1999. Patterns of single-nucleotide polymorphisms in candidate genes for blood-pressure homeostasis. *Nat. Genet.* 22:239–247.

[40] Hey J, Kliman R. 2002. Interactions between natural selection, recombination and gene density in the genes of Drosophila. *Genetics* 160:595–608.

[41] Jordan I, Kondrashov F, Rogozin I, Tatusov R, Wolf Y, Koonin E. 2001. Constant relative rate of protein evolution and detection of functional diversification among bacterial, archaeal and eukaryotic proteins. *Genome Biol.* 2:RESEARCH0053.

[42] Jordan I, Rogozin I, Wolf Y, Koonin E. 2002. Microevolutionary genomics of bacteria. *Theor. Popul. Biol.* 61:435–447.

[43] Jukes T, Cantor C. 1969. *Evolution of protein molecules*, pp. 21–132. New York: Academic Press.

[44] Kimura M. 1957. Some problems of stochastic processes in genetics. *Ann. Math Stat.* 28:882–901.

[45] Kimura M. 1968. Evolutionary rate at the molecular level. *Nature* 217:624–626.

[46] Kimura M. 1969. The number of heterozygous nucleotide sites maintained in a finite population due to steady flux of mutations. *Genetics* 61:893–903.

[47] Kimura M. 1983. *The Neutral Theory of Molecular Evolution.* Cambridge: Cambridge University Press.

[48] Kimura M, Ota T. 1969. The average number of generations until extinction of an individual mutant gene in a finite population. *Genetics* 63:701–709.

[49] King J, Jukes T. 1969. Non-Darwinian evolution. *Science* 164:788–798.

[50] Kondrashov A. 2003. Direct estimates of human per nucleotide mutation rates at 20 loci causing mendelian diseases. *Hum. Mutat.* 21:12–27.

[51] Kondrashov A, Sunyaev S, Kondrashov F. 2002. Dobzhansky-Muller incompatibilities in protein evolution. *Proc. Natl. Acad. Sci. USA* 99:14878–14883.

[52] Kondrashov F, Rogozin I, Wolf Y, Koonin E. 2002. Selection in the evolution of gene duplications. *Genome Biol.* 3:RESEARCH0008.

[53] Kumar S, Hedges S. 1998. A molecular timescale for vertebrate evolution. *Nature* 392:917–920.

[54] Lewontin R, Hubby J. 1966. A molecular approach to the study of genic heterozygosity in natural populations. II. Amount of variation and degree of heterozygosity in natural populations of Drosophila pseudoobscura. *Genetics* 54:595–609.

[55] Li W. 1997. *Molecular Evolution*. Sunderland, MA: Sinauer Associates.

[56] Liberles D, Schreiber D, Govindarajan S, Chamberlin S, Benner S. 2001. The Adaptive Evolution Database (TAED). *Genome Biol.* 2:PREPRINT0003.

[57] Lynch M, Force A. 2000. The probability of duplicate gene preservation by subfunctionalization. *Genetics* 154:459–473.

[58] Makalowski W, Boguski M. 1998. Evolutionary parameters of the transcribed mammalian genome: an analysis of 2,820 orthologous rodent and human sequences. *Proc. Natl. Acad. Sci. USA* 95:9407–9412.

[59] McDonald J, Kreitman M. 1991. Adaptive protein evolution at the Adh locus in Drosophila. *Nature* 351:652–654.

[60] McVean G, Vieira J. 2001. Inferring parameters of mutation, selection and demography from patterns of synonymous site evolution in Drosophila. *Genetics* 157:245–257.

[61] Messier W, Stewart C. 1997. Episodic adaptive evolution of primate lysozymes. *Nature* 385:151–154.

[62] Nekrutenko A, Makova K, Li W. 2002. The K(A)/K(S) ratio test for assessing the protein-coding potential of genomic regions: an empirical and simulation study. *Genome Res.* 12:198–202.

[63] Ng P, Henikoff S. 2002. Accounting for human polymorphisms predicted to affect protein function. *Genome Res.* 12:436–446.

[64] Nielsen R, Yang Z. 1998. Likelihood models for detecting positively selected amino acid sites and applications to the HIV-1 envelope gene. *Genetics* 148:929–936.

[65] Ohno S. 1970. *Evolution by Gene Duplication*. Berlin: Springer-Verlag.

[66] Ohta T. 1973. Slightly deleterious mutant substitutions in evolution. *Nature* 246:96–98.

[67] Ohta T. 1975. Statistical analysis of Drosophila and human protein polymorphism. *Proc. Natl. Acad. Sci. USA* 72:3194–3196.

[68] Ohta T. 1995. Synonymous and nonsynonymous substitutions in mammalian genes and the nearly neutral theory. *J. Mol. Evol.* 40:56–63.

[69] Ohta T. 1997. Role of random genetic drift in the evolution of interactive systems. *J. Mol. Evol.* 44 Suppl 1:S9–14.

[70] Piganeau G, Mouchiroud D, Duret L, Gautier C. 2002. Expected relationship between the silent substitution rate and the GC content: implications for the evolution of isochores. *J. Mol. Evol.* 54:129–133.

[71] Pritchard J. 2001. Are rare variants responsible for susceptibility to complex diseases? *Am. J. Hum. Genet.* 69:124–137.

[72] Reich D, Cargill M, Bolk S, Ireland J, Sabeti P, Richter D, Lavery T, Kouyoumjian R, Farhadian S, Ward R, Lander E. 2001. Linkage disequilibrium in the human genome. *Nature* 411:199–204.

[73] Sawyer S, Dykhuizen D, Hartl D. 1987. Confidence interval for the number of selectively neutral amino acid polymorphisms. *Proc. Natl. Acad. Sci. USA* 84:6225–6228.

[74] Silva J, Kondrashov A. 2002. Patterns in spontaneous mutation revealed by human-baboon sequence comparison. *Trends Genet.* 18:544–547.

[75] Slatkin M. 2000. Allele age and a test for selection on rare alleles. *Philos. Trans. R. Soc. Lond. B Biol. Sci.* 355:1663–1668.

[76] Smith N, Eyre-Walker A. 2002. Adaptive protein evolution in Drosophila. *Nature* 415:1022–1024.

[77] Sunyaev S, Ramensky V, Koch I, Lathe Wr, Kondrashov A, Bork P. 2001. Prediction of deleterious human alleles. *Hum. Mol. Genet.* 10:591–597.

[78] Suzuki Y, Gojobori T. 1999. A method for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 16:1315–1328.

[79] Suzuki Y, Nei M. 2001. Reliabilities of parsimony-based and likelihood-based methods for detecting positive selection at single amino acid sites. *Mol. Biol. Evol.* 18:2179–2185.

[80] Suzuki Y, Nei M. 2002. Simulation study of the reliability and robustness of the statistical methods for detecting positive selection at single amino Acid sites. *Mol. Biol. Evol.* 19:1865–1869.

[81] Takahata N. 1991. Statistical models of the overdispersed molecular clock. *Theor. Popul. Biol.* 39:329–344.

[82] Takahata N, Satta Y, Klein J. 1995. Divergence time and population size in the lineage leading to modern humans. *Theor. Popul. Biol.* 48:198–221.

[83] Takano T. 1998. Rate variation of DNA sequence evolution in the Drosophila lineages. *Genetics* 149:959–970.

[84] Templeton A. 1996. Contingency tests of neutrality using intra/interspecific gene trees: the rejection of neutrality for the evolution of the mitochondrial cytochrome oxidase II gene in the hominoid primates. *Genetics* 144:1263–1270.

[85] Ting C, Tsaur S, Wu C. 2000. The phylogeny of closely related species as revealed by the genealogy of a speciation gene, Odysseus. *Proc. Natl. Acad. Sci. USA* 97:5313–5316.

[86] Ting C, Tsaur S, Wu M, Wu C. 1998. A rapidly evolving homeobox at the site of a hybrid sterility gene. *Science* 282:1501–1504.

[87] Tsaur S, Wu C. 1997. Positive selection and the molecular evolution of a gene of male reproduction, Acp26Aa of Drosophila. *Mol. Biol. Evol.* 14:544–549.

[88] Urrutia A, Hurst L. 2001. Codon usage bias covaries with expression breadth and the rate of synonymous evolution in humans, but this is not evidence for selection. *Genetics* 159:1191–1199.

[89] Waterston R, et a. 2002. Initial sequencing and comparative analysis of the mouse genome. *Nature* 420:520–562.

[90] Watterson G. 1987. Estimating the proportion of neutral mutants. *Genet. Res.* 50:155–163.

[91] Weinreich D, Rand D. 2000. Contrasting patterns of nonneutral evolution in proteins encoded in nuclear and mitochondrial genomes. *Genetics* 156:385–399.

[92] Wright S. 1938. The distribution of gene frequencies under irreversible mutations. *Proc. Natl. Acad. Sci. USA* 24:253–259.

[93] Wu C, Li W. 1985. Evidence for higher rates of nucleotide substitution in rodents than in man. *Proc. Natl. Acad. Sci. USA* 82:1741–1745.

[94] Wu C, Palopoli M. 1994. Genetics of postmating reproductive isolation in animals. *Annu. Rev. Genet.* 28:283–308.

[95] Wyckoff G, Wang W, Wu C. 2000. Rapid evolution of male reproductive genes in the descent of man. *Nature* 403:304–309.

[96] Yang H, Tanikawa A, Kondrashov A. 2001. Molecular nature of 11 spontaneous de novo mutations in Drosophila melanogaster. *Genetics* 157:1285–1292.

[97] Yang Z. 1998. Likelihood ratio tests for detecting positive selection and application to primate lysozyme evolution. *Mol. Biol. Evol.* 15:568–573.

[98] Yang Z, Bielawski J. 2000. Statistical methods for detecting molecular adaptation. *Trends Ecol. Evol.* 15:496–503.

[99] Yang Z, Nielsen R. 1998. Synonymous and nonsynonymous rate variation in nuclear genes of mammals. *J. Mol. Evol.* 46:409–418.

[100] Yang Z, Nielsen R. 2000. Estimating synonymous and nonsynonymous substitution rates under realistic evolutionary models. *Mol. Biol. Evol.* 17:32–43.

[101] Yang Z, Nielsen R. 2002. Codon-substitution models for detecting molecular adaptation at individual sites along specific lineages. *Mol. Biol. Evol.* 19:908–917.

[102] Yang Z, Nielsen R, Goldman N, Pedersen A. 2000. Codon-substitution models for heterogeneous selection pressure at amino acid sites. *Genetics* 155:431–449.

[103] Yu N, Fu Y, Li W. 2002. DNA polymorphism in a worldwide sample of human x chromosomes. *Mol. Biol. Evol.* 19:2131–2141.

[104] Zhang J, Rosenberg H, Nei M. 1998. Positive Darwinian selection after gene duplication in primate ribonuclease genes. *Proc. Natl. Acad. Sci. USA* 95:3708–3713.

[105] Zuckerlandl E, Pauling L. 1965. *Evolutionary divergence and convergence in proteins*, pp. 97–166. New York: Academic Press.

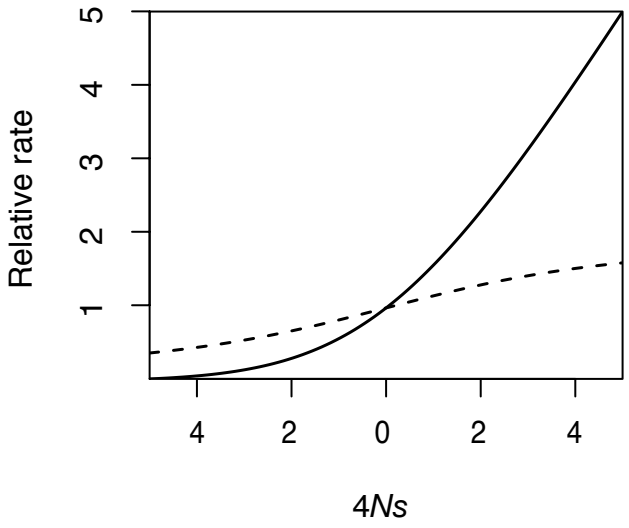

Figure 1: Relative rate of selected to neutral substitutions, $\frac{4Ns}{1-e^{4Ns}}$ (solid), and heterozygosity, $\frac{2(4Ns-1+e^{-4Ns})}{4Ns(1-e^{-4Ns})}$ (dashed) as a function of $4Ns$ [48].

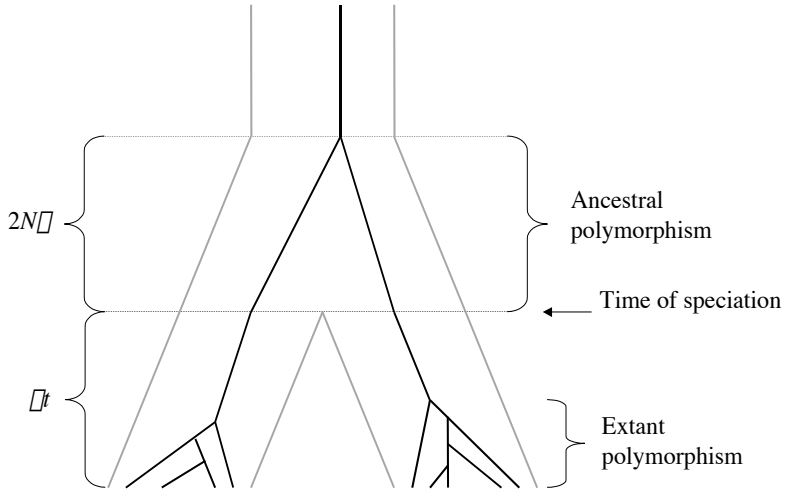

Figure 2: Genealogy reflecting the proportion of sequence divergence due to segregation of ancestral polymorphism, $4Nu$, and divergence after speciation, $2\mu t$. The grey lines indicate the split of one species into two and the black lines represent a single genealogy, consisting of polymorphism in both the extant species, divergence since the time of speciation and ancestral polymorphism present at the time of speciation.

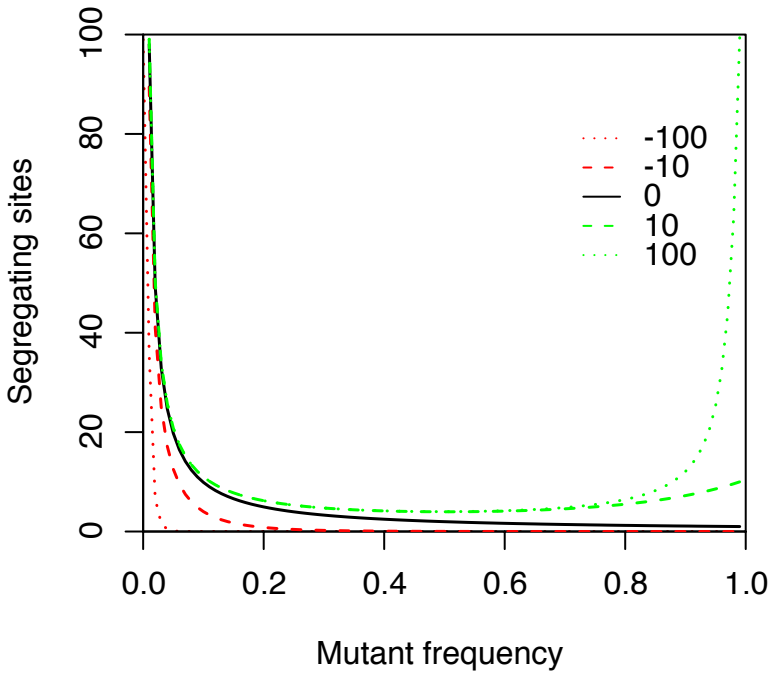

Figure 3: Frequency spectrum for sites under positive, $4Ns > 1$ (green), negative, $4Ns < -1$ (red) and no selection $4Ns = 0$. The frequency spectrum or expected number of mutations in a population as a function of their frequency is given by $\phi(x) = \frac{4N\mu}{x(1-x)}\left(\frac{1-e^{4Ns(1-x)}}{1-e^{-4Ns}}\right)$ [23] [92].